Image Manipulation Detection by Multi-View Multi-Scale Supervision

-Supplementary Materials-

Xinru Chen[1,2*], Chengbo Dong[1,2*], Jiaqi Ji[1,2], Juan Cao[3,4], Xirong Li[1,2†]
[1]MoE Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China
[2]AIMC Lab, School of Information, Renmin University of China
[3]Institute of Computing Technology, Chinese Academy of Sciences
[4]State Key Laboratory of Media Convergence Production Technology and Systems

This supplementary material provides details that could not be included in the paper submission due to space limitations. Additional experimental results are list as follows.

- Backbone comparison in Section 1, trained on DEFACTO-84k and evaluated on DEFACTO-12k;
- Fusion strategy in Section 2, trained on CASIAv2 and evaluated on CASIAv1;
- Robustness evaluation in Section 3.
- Hyper-parameter sensitivity evaluation in Section 4, trained on CASIAv2 and evaluated on CASIAv1;
- Qualitative results with SOTA works including failure analysis in Section 5;

1. Comparison on Backbone

We provide the performance of distinct semantic segmentation networks in Table 1, showing the improved FCN used in the current work is the best.

Segmentation network	Pixel-F1	Image-F1	Image-AUC	Com-F1
Improved FCN-16	**0.546**	**0.709**	**0.840**	**0.617**
FCN-16	0.337	0.699	0.774	0.455
U-Net	0.132	0.517	0.540	0.210
DeepLabV3	0.249	0.526	0.645	0.338
DeepLabV3+	0.279	0.509	0.651	0.360

Table 1. **Performance of distinct segmentation networks**, all trained following Setup#0 (**L650**), *i.e.* the segmentation loss only.

2. Fusion Strategy

We claim novelty for the use of DA, as previously, branch fusion was implemented by bilinear pooling (BP) [32]. We compare DA and BP within our MVSS-Net framework. As Table 2 shows, DA is more effective than BP.

Branch Fusion	Pixel-F1	Image-F1	Image-AUC	Com-F1
DA	**0.538**	**0.799**	**0.886**	**0.643**
BP	0.424	0.772	0.845	0.547

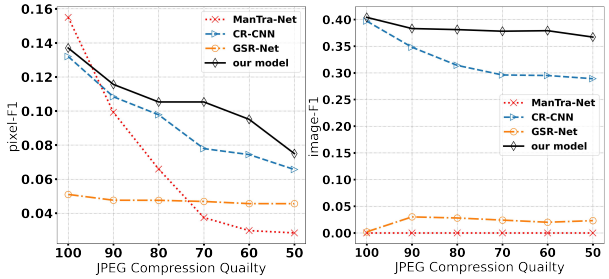
Table 2. **Performance of two MVSS-Nets, one uses DA and the other uses BP for branch fusion**.

3. Robustness Evaluation

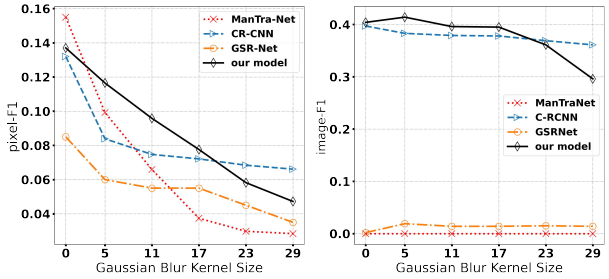
JPEG compression and Gaussian blur are separately applied on each test image in DEFACTO-12k. Performance curves in Fig. 1 show better robustness of MVSS-Netwhen compared with the baselines.

We show in Fig. 3 manipulation segmentation of some test images under a decreasing level of JPEG compression quality. Fig. 4 shows manipulation segmentation given an increasing level of Gaussian blur.

*Xinru Chen and Chengbo Dong contribute equally to this work.
†Corresponding author: Xirong Li



(a) Performance curves w.r.t. JPEG compression



(b) Performance curves w.r.t. Gaussian blurs

Figure 1. **Robustness evaluation against (a) JPEG compression and (b) Gaussian blurs**. Test set: DEFACTO-12k.

4. Hyper-parameter Sensitivity Evaluation

For the two hyper-parameters, α and β, that balance $loss_{seg}$, $loss_{edg}$ and $loss_{clf}$ in the joint loss, we set $(\alpha, \beta) = (0.16, 0.04)$ according to our ablation study conducted on DEFACTO. The same values are used when we train MVSS-Net on CASIAv2 to compare with the state-of-the-arts.

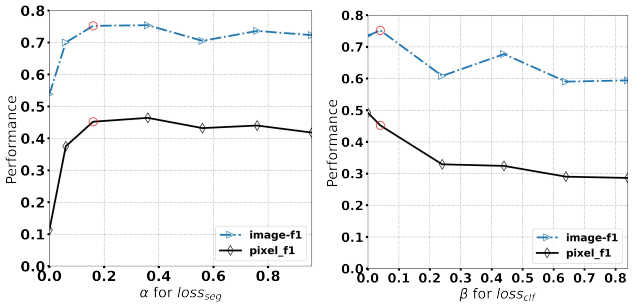
$$Loss = \alpha \cdot loss_{seg} + \beta \cdot loss_{clf} + (1-\alpha-\beta) \cdot loss_{edg}$$

In order to evaluate how sensitive MVSS-Net is w.r.t. the two hyper-parameters, we conduct the following two studies:

- Fix β as 0.04 and vary α in $[0, 1-\beta]$;
- Fix α as 0.16 and vary β in $[0, 1-\alpha]$.

Models are trained fully on CASIAv2 and tested on CASIAv1. The default decision threshold of 0.5 is used for both pixel-level and image-level binary classification.

As the performance curves in Fig. 2(a) show, MVSS-Net maintains a good performance given a relatively wide range of α, indicating that α is not highly sensitive. Fig. 2(b) shows that the overall performance degenerates as β increases. This result suggests that a proper balance between model sensitivity and specificity cannot be simply achieved by tuning the influence of $loss_{clf}$, and novel designs in the network (as we articulate in the paper) are necessary. As the setting of $(\alpha, \beta) = (0.16, 0.04)$ performs well on two training sets and five test sets, we recommend it as a default choice for using MVSS-Net.



(a) $\beta = 0.04$ (b) $\alpha = 0.16$

Figure 2. **Hyper-parameters sensitivity evaluation of MVSS-Net**. Red circles denote $(\alpha, \beta) = (0.16, 0.04)$ used throughout the paper.

5. More Qualitative Comparison with State-of-the-art

Fig. 5 presents qualitative results between MVSS-Net and three open-sourced state-of-the-art models, *i.e.* ManTra-Net, GSR-Net and CR-CNN. Compared to the existing models, MVSS-Net is more sensitive to the manipulation areas, while at the same being highly specific, giving zero response on the authentic images in the last row rows.

Given the challenging nature of the task, failures are inevitable, see Fig. 6. The test image in the first row is a copy-move manipulated image from the COVER dataset, showing little boundary artifact between tampered area and the rest. The second image (splicing), selected from DEFACTO-12k, has a tiny tampered area that none of the current models can detect it with success. We conjecture that a pre-selection of a region of interest may remedy such a situation. Indeed, given a manually cropped region as a new input, see Fig. 7, the prediction of MVSS-Net hits the tampered region, *i.e.* the clock. One more failure is shown in the third row, which is an inpainting from NIST16. MVSS-Net succeeded in discriminating differences between tampered and authentic regions, see its segmentation map at the last row, yet failed to recognize the background which was actually manipulated.

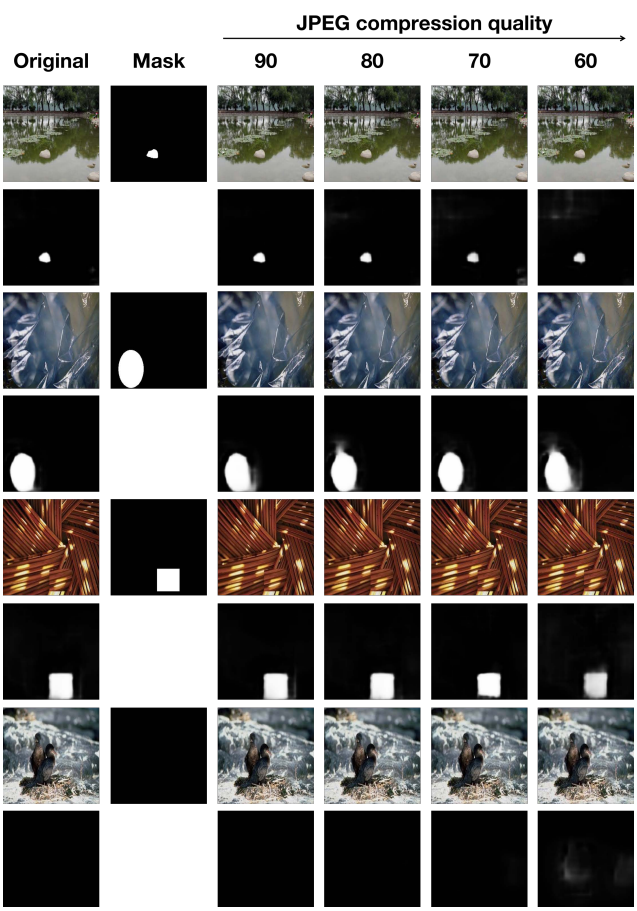


Figure 3. **Pixel-level manipulation detection results under a decreasing level of JPEG compression quality**. Below each image is the corresponding manipulation segmentation predicted by MVSS-Net. Even at a relatively low quality of 60 (the last column), MVSS-Net still performs well.

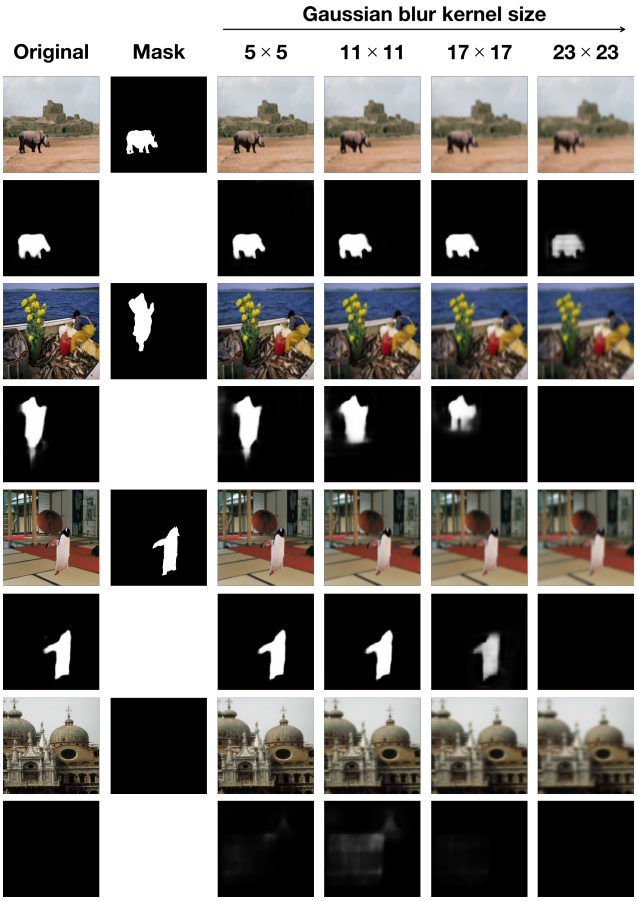


Figure 4. **Pixel-level manipulation detection results given an increasing intensity of Gaussian blurs**. Below each image is the corresponding manipulation segmentation predicted by MVSS-Net. Even when the test images are smoothed with a large kernel of 17×17, MVSS-Net still makes reasonable predictions, see the second last column.

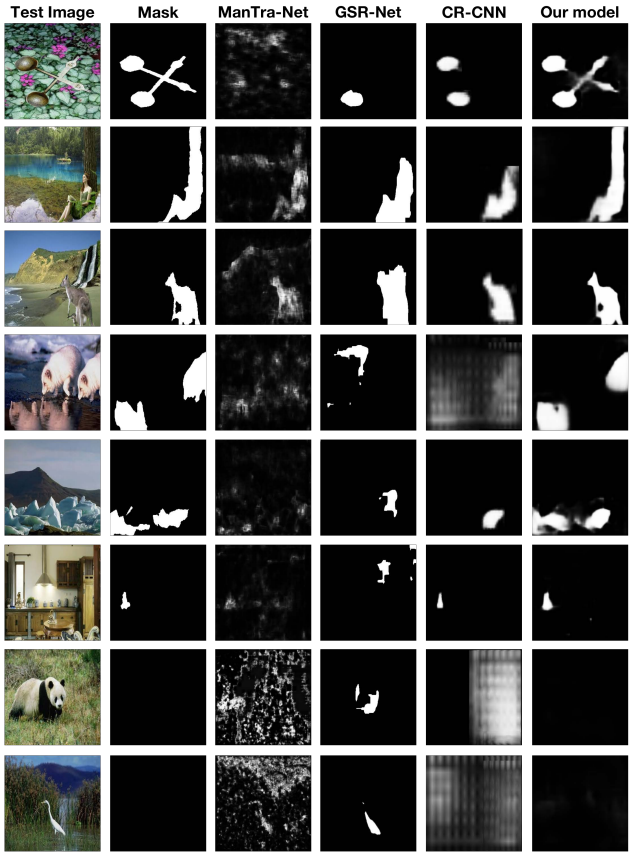


Figure 5. **Qualitative comparison between MVSS-Net and the state-of-the-art**. All test images are selected from CASIAv1, with manipulated images given in the top six rows and authentic images in the bottom two rows.

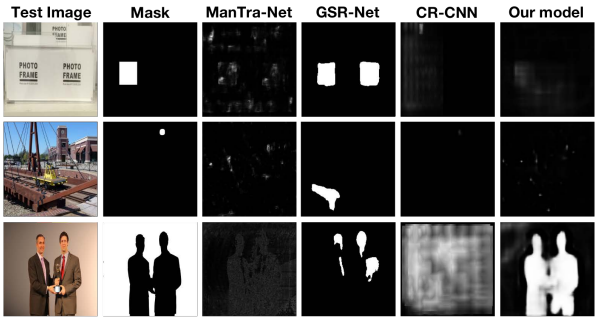


Figure 6. **Failure cases**. Manipulation from the top to bottom is copy-move, splicing and inpainting, with test images chosen from COVER, DEFACTO-12k and NIST16, respectively.

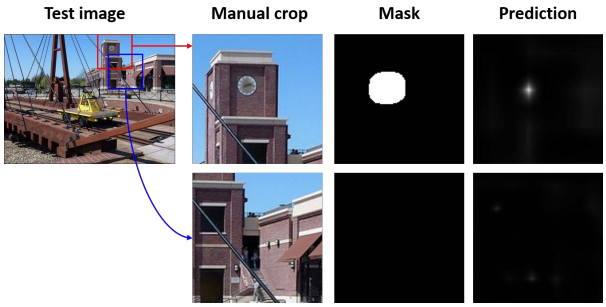


Figure 7. **Segmentation on manually cropped regions over tiny manipulation and authentic area**. Pre-selection can remedy false negatives of MVSS-Net on tiny manipulation as presented in the first row. Meanwhile, it gives proper prediction on the authentic patch in the second row.